DOCUMENT RESUME

ED 188 141 CS 005 507

AUTHOR Fitzgerald, Thomas P.; Fleming, James T.
TITLE Test-Wiseness in Competency Testing Programs.
PUB DATE Mar 80
NOTE 15p.; Paper presented at the Annual Meeting of the International Reading Association Eastern Regional Conference (3rd, Niagara Falls, NY, March 27-29, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Intermediate Grades; *Language Patterns; *Minimum Competency Testing; *Reading Research; *Reading Tests; Syntax; *Test Construction; Test Items; *Test Wiseness; Vocabulary

ABSTRACT

In a study concerned with the function of test wiseness in competency testing and with the impact of question stem structure within a reading competency test, 43 fourth grade students and 38 sixth grade students were administered one form each of the experimental material. This material consisted of ten questions presented in the stylized language pattern of the original format (drawn from the Basic Competency Test in reading--1975) and ten questions recast by changing the vocabulary, the syntax, or both. Data were analyzed by total score, stylized question stems, and altered question stems for grade, reading, and performance levels. Results showed that (1) sixth grade students recorded higher total scores, (2) fourth grade students achieved higher scores with the altered (less formal) stems, and (3) questions reflecting syntactic restructuring resulted in the highest percentage of correct responses when compared to vocabulary changes or a combination of alterations. Data indicate that care should be taken when structuring test items to assure direct and clear syntactic arrangement of question stems, and that students will score according to their ability when question stems more closely parrallel their own language patterns. (JM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED188141

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TEST-WISENESS IN COMPETENCY TESTING PROGRAMS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Thomas P. Fitz-gerald

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

THOMAS P. FITZGERALD
NEW YORK STATE EDUCATION DEPARTMENT

JAMES T. FLEMING
STATE UNIVERSITY OF NEW YORK AT ALBANY

CS 005 507

Test-wiseness in Competency Testing Programs

Thomas P. Fitzgerald
New York State Education Department

James T. Fleming
State University of New York at Albany

Introduction

Major educational decisions are now based on the results of various state and local competency tests. The influence of students' test-wiseness has not been systematically addressed in relationship to competency testing. However, the general issue of test-wiseness has received much attention in recent years. Millman, Bishop and Ebel (1965) catalogued types of test-wiseness and defined this trait as the capacity of students to use the characteristics and format of a test to achieve a higher score. In the analysis of the characteristics of test-wiseness, they identified the familiar or stereotyped phraseology as one element potentially influencing performance. Since the target population for competency testing programs are those functioning at the lower end of the achievement scale, test-wiseness in competency testing assumes major importance. Certain students may not be producing scores representative of their abilities because they are unable to use the characteristics and format of the test to their advantage.

Millman et al (1965) categorized the elements of test-wiseness as two major types: those elements which are independent of the test maker or test purpose; and those dependent on the maker and purpose. For a competency testing program, the dependent elements appear more critical and will be the concern of this paper. Millman et al (1965) analyzed the intent and cue-using strategies as part of the dependent elements of test-wiseness.

One cueing strategy influencing a reader's ability to respond correctly to

an item stem is the syntactic arrangement of that stem. Denner (1970) reported data indicating that problem readers were less competent in synthesizing whole sentences presented in a linear form when compared to competent readers. In a related area, Gronlund investigated the effect of vocabulary and sentence structure on performance. The data from this study were interpreted as indicating that the more difficult sentence structures, sentence ambiguities and vocabulary tend to divert the attention of the reader away from the subject matter of the test, thereby lowering scores.

A second cueing strategy influencing test-wiseness is the relationship between the oral language pattern of the test-taker and the written language pattern of the test. Nolen (1972) found higher performance levels when non-standard dialect material was substituted in one standardized test. Knapp (1975) also suggested classroom instructional and testing techniques for using dialect differences to an advantage. Since the educational and testing impact of vernacular language has not been fully explored, a more salient cueing element is sentence structures used in testing instruments.

Objective test formats have frequently been constructed using a formal language pattern in an effort to avoid flawed test items (N.Y.S. Education Department, 1975). Even with such pie caution in test construction, some writers have continued to advocate teaching attending behavior and a set of response procedures for particular question types (Erichson, 1972). The multiple choice format appears particularly susceptable to test-wiseness variations. Rowley (1974) reported that the multiple choice format was more advantageous to thise students who were "highly testwise and willing to take risks." In addition, Strang (1977) confirmed the negative effects of technical and unfamiliar options when using a multiple choice test format.

When considering the potential influence of test-wiseness strategies such as language patterns, the question arises as to whether this construct has an equal and a constant impact on scores across grade levels. Sarnacki (1979)

described two possible theoretical perspectives for test-wiseness. First, these elements may be viewed as an additional source of variance in test scores, due mainly to flawed test items. A second viewpoint is that test-wiseness may be a persistent attribute of the individual test-taker. Some empirical evidence has been reported to support the consistent longitudinal impact of these factors (Crehan, Koehler and Slakter, 1974). Many states have recently initiated some form of a competency testing program (Pipho, 1977). The data produced by such tests are used to determine promotion, course selection, and finally graduation. The literature is resplendent with argumentation both for and against the concepts and the applications of competency testing. Since these tests are now in use, the question of test-wiseness becomes a vital concerm and the focus of this study.

Problem

This study investigated the impact of question stem structure within a reading competency test. The purpose of the study was to analyze the test-wiseness factor of language structure used in question stems and its effect on fourth and sixth grade students. Would transformation of question stems into a less formal, more direct phrasing produce higher scores? The results were analyzed by grade and reading levels to determine any differential or longitudinal impact based on the language structure of question stems. In essence, this paper investigates the impact of formal and less formal language patterns in question stems within a reading competency test.

Methodology

Subjects for this study were drawn from an urban school in up-state New York. Two intact classes of fourth grade students (N=43) and two classes of sixth grade students (N=38) were administered one form of the experimental material. The testing took place within a classroom setting and was adminstered by the authors

using a uniform set of directions. Although no time limitations were imposed, the majority of students completed the testing within thirty-five minutes. The teachers of these four classes recorded a reading level of each student based on standardized test scores. The test data of six students were eliminated before analysis because of incomplete testing or low reading ability.

The material used in this experiment was drawn from one form of the Basic Competency Test (BCT) in reading (1975). This test was based on Sharon's (1973-74) tabulations of what types of material read by adults and how frequently the items were read. The BCT was designed as a predictive instrument to determine which intermediate students possessed reading skills necessary to function as an adult in society. The multiple choice format used in the BCT required the reader to select the most appropriate answer from among five choices following the reading of a stimulus item. The test was constructed to evaluate mainly literal comprehension and, as such, most answers were explicitly stated within the stimuli.

For the experimental material, twenty questions based on fifteen stimuli were selected for this study. Each question was maintained in its original form and also recast into one of three conditions. The new question stems reflected a recasting by either changing the vocabulary (paraphrasing), syntax, or a combination of these conditions.

Two experimental forms were arranged which counter-balanced the original, stylized stems with the recast stem in the second form. Each student responded to ten questions using stylized language pattern and ten reflecting the altered language patterns. Data from the study indicated that there were non-significant differences between the two forms generated comparing the performance of fourth grade students ($F=.365$) and of sixth grade students ($F=.017$).

Both forms retained the same multiple choice format of the original BCT. The questions were displayed at the top of an individual sheet with the stimulus printed below. The students were required to identify the appropriate answer and to record that choice on a separate answer sheet.

The scoring of the experimental material produced a total score for the test form, a score for the stylized questions and a score for the altered stems. The data were analyzed by grade level, reading level and type of stem alteration.

Findings and Results

Table 1 displays the scoring pattern for both grades based on the twenty questions of each form. As expected, higher performances were produced by the sixth grade subjects on both forms ($F=51.71$, $p<.001$). The overall performances of sixth graders represented a 71% level of accuracy and 46% level of accuracy for fourth graders. The lower performances by fourth graders reflected the readability levels in the stimuli of experimental materials which ranged from fourth to eight grade levels.

Student performance with the two arrangements of question stems are also displayed in Table 1. The fourth grade subjects recorded higher performances with the altered stems while the sixth grade subjects were more successful with stylized stems. Neither score differences, however, were statistically dignificant ($F=.974$; $F=.537$). In general, these data indicate that recasting question stems produce small differences in test performance.

Table 2 arranges the correlation data between total score, reading level, stylized stem score and altered stem score. All the correlations reach levels of significance except between reading levels and stylized stem performance with fourth grade students. These data indicated that for fourth graders higher performance rates more closely paralleled higher reading levels with the altered stems (.3054) and, to a lesser degree, with the stylized stems (.1954). The reverse was indicated with sixth grade subjects in that performance with stylized stems were highly related to reading level (.7054). The data again indicated that at the sixth grade level the two question forms were highly correlated (.6187) and at a lower rate by fourth graders (.3343) when relating the two question formats. In addition, the correlational data indicate differences in performances between total scores and reading levels. At the sixth grade level,

student's total score performance more closely reflect reading levels than at fourth grade level.

The data were further analyzed to determine whether performance differences with the question types were evident when high and low performers were identified. Individual students were categorized as either high performers or low performers on this test based on the relationship between their source and the median score for all subjects. Because of this grouping, the sixth grade subjects have a larger number of high performer

Table 3 displays a comparison of subjects who had a higher performance with either the stylized or altered form of question stems. At the fourth grade level, nine students were more successful with the altered stems compared to five high performers who were more successful with the stylized stems. A reverse in success pattern occurred at the sixth grade level for high performers. With low performers at both levels, the data indicate no differences in success with either question type.

The twenty questions used in this study were recast using one of three conditions that described the principal change between the original stem and the altered stem. Six alterations reflected a paraphrasing of vocabulary. An additional six stems had a change in syntax where words were moved or deleted. The remaining eight questions had both a change in vocabulary and syntax. The data on percentage of correct responses by type of recasting is represented in Table 4 for both grade and reading levels. High and low reading levels reflect student performance on a standardized reading test (California Test of Basic Skills) as either above or below the median score for that grade.

For both grades, the high achievement students produced higher scores with the syntactically recast questions. For low achieving students, the data also indicated a similar performance when there was a syntactic change in the stem. For both grade and ability levels, the lowest performance was recorded with vocabulary changes. In general, the data appeared to support the need for directness and clarity in syntax if students are to achieve according to their ability.

Conclusions

This study focused on the function of test-wiseness in competency testing. Question stems from the Basic Competency Test were recast by either changing the vocabulary level, syntax or both vocabulary and syntax. Students in grades four and six were administered either of the two forms which had the original stems and altered stems counter-balanced. The data were analyzed by total score, by formed or stylized, and, by altered or less formal question stems for grade, reading and performance levels.

As expected, the sixth grade subjects recorded higher total scores. However, the fourth grade subjects achieved higher scores with the altered or less formal stems. The longitudinal test-taking strategies identified by Gronlund (1976) are supported by these data in that the older students were more adept at handling the stylized or more formal question format. The fourth grade subjects, especially those at the higher reading achievement levels, were more successful with the altered stems, possibly recognizing them as closer approximately to their own oral language patterns. It is possible that what the sixth graders are demonstrating is a greater sensitivity to more formal test language patterns as a result of greater exposure to them in a school setting. In addition, the recast stems did not have uniform appeal for all the students. Those which reflected syntactic restructuring resulted in the highest percentage of correct responses when compared to changes in the vocabulary or a combination of alterations.

The factors of sentence structure and vocabulary have been identified as part of test-wiseness and the data from this study reveal non-significant differences in student performances with stylized and altered question stems from a competency testing program. The data did indicate that care should be taken when constructing test items to assure direct and clear syntactic arrangement of question stems. In addition, students will score according to their ability when question stems more closely parallel their oral language patterns--in this instance, a less formal style.

References

Basic Competency Test in Reading. New York State Education Department. Albany, New York, 1975.

Crehan, K., Koehler, R., and Slakter, M. Longitudinal studies of test-wiseness. Journal of Educational Measurement, 11, 1974, 209-212.

Denner, B. Representational and syntactic competence of problem readers. Child Development, 41, 1970, 881-887.

Erickson, M. Test sophistication. Journal of Reading, 16, #2, 1972, 140-144

Gronlund, N. Measurement and Evaluation in Teaching (3rd edition) New York, Macmillan, 1976.

Improving the Classroom Test. New York State Education Department. Albany, New York, 1975.

Knapp, M. Black dialect and reading: What teachers need to know. Journal of Reading, 19, #3, 1975, 231-236.

Millman, J., Bishop C., and Ebel, R. An analysis of test-wiseness. Educational and Psychological Measurement, 25, 1965, 707-726.

Nolen, P. Reading of non-standard dialect materals: A study at grades two and four. Child development, 43, 1972, 1092-1097.

Pipho, C. Minimal competency testing: A look at state standards. Educational Leadership, 34, #7, 1977, 516-520.

Rowley, G. Which examinees are most favored by the use of multiple choice tests? Journal of Educational Measurement, 11, #1, 1974, 15-23.

Sarnacki, R. An examination of test-wiseness in the cognitive test domain. Review of Educational Research, 49, #2, 1979, 252-279.

Sharon, A. What do adults Read. Reading Research Quarterly. 9, #2, 1973-74, 148-168.

Strang, H. The effects of technical and unfamiliar options on guessing on multiple choice test items. Journal of Educational Measurement, 14, #3, 1977, 253-259

Figure 1

SELECTED, STYLIZED AND ALTERED QUESTIONS

1. If you were hired as a pipefitter, how much an hour would you be paid?

   What are pipefitters paid?

2. Normally, for how many days in a row should you apply the medicine? How many days do you put the medicine on?

3. According to the instructions, how much bleach should be used in a gallon of water for cleaning kitchens?

   To clean kitchens, how much bleach goes in a gallon of water?

4. What is the possible fine for not telling the truth when you apply for a drivers license?

   What is the fine for lying on a drivers license?

5. For how many days would your license be suspended if you were convicted fo driving at a speed of 65 miles per hour in a 50 mile per hour zone?

   If you drive 65 in a 50 miles an hour zone, they would take your license away for

6. At which restaurant could you eat if you wanted to have dinner between 5:30 PM and 6:00 PM?

   Dinner between 5:30 PM and 6:00 PM is at

7. How much would it cost to take a two-year subscription to ESSENCE?

   How much does it cost to get ESSENCE for two years?

8. To whom should union members talk if they feel the union is doing wrong?

   Who should you talk to in the union?

TABLE 1

Comparison of score performance across Forms for the fourth (N=43) and sixth (N=38) grade subjects.

		Scores and (SD)		
Grade	Form	Stylized	Altered	Total
4th	A	4.63(2.01)	4.77(1.97)	4.70
	B	4.47(1.93)	4.57(2.02)	4.52
	A + B	4.55(1.95)	4.67(1.99)	4.61
6th	A	7.73(2.32)	6.63(2.22)	7.16
	B	6.94(2.04)	7.52(2.22)	7.23
	A + B	7.34(2.19)	7.07(2.24)	7.19

Table 2

Pearson's Product Moment correlation between stylized, altered, total scores for fourth ad sixth grade subjects by reading levels.

	Total Score		Stylized Score		Altered Score	
	4th	6th	4th	6th	4th	6th
Reading Levels	.2594*	.6697***	.1954	.7054***	.3054*	.5710*
Altered	.7101***	.8615***	.3343*	.6187**		
Stylized	.5704**	.8825***				

* = significant at .05 level
** = significant at .01 level
*** = significant at .001 level

Table 3

Comparison between high and low performance across question types for fourth (N=43) and sixth (N=38) grade subjects.

Grade	Performance level		Score differences by question types: Stylized	Altered	No Different
4th	high (N=19)		5	9	5
	low (N=24)		10	10	4
		Total	15	19	9
6th	high (N=26)		14	8	4
	low (N=12)		5	5	2
		Total	19	13	6

Table 4

Percent of correct responses by type of altered question stem across grade and reading levels.

Question Stem Alteration	Percent of Correct Response			
	Fourth Grade		Sixth Grade	
	high	low	high	low
vocabulary	45%	40%	78%	61%
syntax	67%	45%	85%	71%
vocabulary and syntax	41%	45%	80%	60%
Totals	49.6	43.3	79	62.6
	46.7%		70.7%	